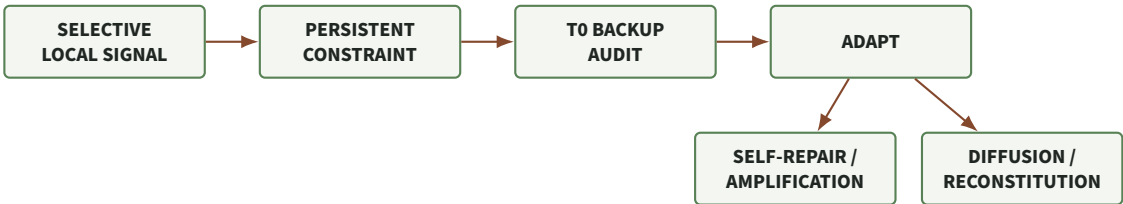


Where Does the Function Go?

A Research Proposal for Functional Reconstitution and Pre-Commitment Recoverability in Language Models

A detectable signal is not automatically the mechanism that produced the behaviour.



Author	Dennis Hedegreen
Publisher	Hedegreen Research
Date	23 September 2026
Issue	MI-001
Version	v0.9 documentation revision
DOI	Pending
Code	GitHub: where-does-the-function-go
Contact	contact@hedegreenresearch.com
Status	Working paper — research proposal / pre-results
Companion	Article 200: Where Does the Function Go?

Abstract

Mechanistic interpretability frequently identifies internal model components whose activity predicts or causally affects a target behaviour. Such localization results are informative, but they do not by themselves establish that the identified component is the unique or necessary implementation of the function. Large language models may contain redundant routes, distributed representations, or mechanisms that reorganize under further optimization.

This proposal develops two connected experimental programmes.

The first tests **functional reconstitution**. Starting from hallucination-associated H-Neurons reported by Gao et al. (2025), the experiment asks whether a behavioural function that is selectively disrupted by persistent neuron-level constraints can recover during continued optimization through a new internal implementation. Version 0.6 strengthens the redundancy test further: candidate backups are measured conditionally at time zero, before adaptation, and later outcomes are separated into pre-existing self-repair, adaptive amplification of a weak backup, distributed redistribution, restricted-subspace compensation, and de novo or near-null reconstitution. The strongest label is reserved for recovery that cannot already be explained by a conditionally causal backup at time zero.

The second programme tests **pre-commitment recoverability**. Rather than asking whether a model “knows” that an answer is wrong, the proposal asks whether hidden state before commitment predicts whether the model’s own continuation distribution still assigns meaningful probability to a verified correction. Natural-support continuation with termination allowed remains primary, but version 0.6 adds a one-token termination-barrier control so stopping policy can be separated from post-stop recoverability. Forced-compute continuation with termination constrained is retained only as a separate secondary intervention. Recoverability is compared against simpler confidence, prompt-only and residual-stream baselines. J-space representations exposed by the Jacobian lens remain one candidate explanatory level, not a privileged assumption.

A later bridge experiment tests whether local H-Neuron-associated mechanisms and distributed J-space-like states are causally coupled. A smaller proposition-substitution task is retained as a controlled model organism for self-monitoring failures. Version 0.7 also adds an optional downstream state-transfer branch inspired by Cache-to-Cache communication: if a source model fails, can a separate receiver extract task-relevant information from the source model’s KV-cache that is not recovered from the same visible text alone? This branch is explicitly separated from the core recoverability claim.

The proposal is intentionally pre-results. It defines targets, controls, decision rules, failure modes and compute gates before substantive adaptation runs are conducted.

Keywords: mechanistic interpretability; large language models; neuron localization; H-Neurons; J-space; causal intervention; model adaptation; recoverability.

Preferred citation

Hedegreen, Dennis. 2026. *Where Does the Function Go? A Research Proposal for Functional Reconstitution and Pre-Commitment Recoverability in Language Models*. Hedegreen Research Working Paper, MI-001, v0.9 documentation revision. DOI pending.

Companion article

Public-facing companion: [Article 200, *Where Does the Function Go?*](#). The article gives the readable research narrative; this working paper carries the current protocol, controls, decision rules and implementation boundaries.

Protocol revision note – v0.9

Version 0.9 adds the public software repository and an implementation-status note. The v0.7 scientific protocol, including the v0.8 figure-layout correction, is unchanged. No empirical result is added.

This release preserves the v0.6 adversarial-review revisions and adds Cache-to-Cache (C2C) as a bounded downstream state-transfer test. The new branch asks whether a failed model's internal KV-cache can improve a separate receiver beyond the same visible text, while treating the cache fuser as a potential confound that must be trained and frozen on disjoint data. C2C is not used as evidence for J-space, recoverability, consciousness, or self-knowledge. No empirical result is added by this revision.

How to read this paper

This is a pre-results research proposal. It does not claim that functional reconstitution occurs, that H-Neurons form a universal hallucination circuit, that recoverability is equivalent to “knowing” an answer is wrong, or that J-space establishes consciousness. The paper specifies experiments, gates, controls and implementation requirements that must survive before stronger claims are available.

Public implementation status

The companion software is available at <https://github.com/DennisHedegreen/where-does-the-function-go>. The initial public release provides provenance checks, source-adaptor contracts, a persistent-mask CPU fixture and conditional-backup audit scaffolding (M0–M3). Its 47 CPU tests validate software behavior on synthetic data and a small gated-MLP model. They do not establish H-Neuron replication, a real Gate 0 pass, CoAx reference replication, or functional reconstitution. Full model execution is planned on RunPod; model/data revisions, runtime settings and statistical methods remain to be frozen. The software retains the v0.8 paper as its scientific baseline.

What this proposal tests

The proposal asks two primary empirical questions. First, if a selectively causal internal locus is kept persistently unavailable while optimization continues, can the affected behaviour return through a new selectively causal implementation? Second, before a model commits to an incorrect or incomplete answer, does its hidden state predict whether bounded additional computation can rescue the trajectory beyond simpler confidence signals? A later bridge experiment tests whether local H-Neuron-associated mechanisms and distributed J-space-like states are causally related.

The Proposal in Two Minutes

The localization problem. A neuron or direction can predict a behaviour without being the unique mechanism that implements it. The proposal separates predictive localization, causal contribution, functional necessity, and implementation identity.

Functional reconstitution. H-Neurons are first replicated and causally tested. If a selective locus survives Gate 0, it is kept persistently unavailable while the rest of the model is allowed to adapt. A strong reconstitution claim requires R0–R4.

Redundancy before learning. Alternative routes are tested conditionally immediately after masking and before adaptation. A strong de novo-reconstitution claim requires that later recovery cannot already be explained by a conditionally causal backup at time zero. Weak backups that grow during learning are classified separately as adaptive amplification.

Recoverability, not “knowing”. The primary question is whether the model’s own continuation distribution, from the exact pre-stop prefix with termination still allowed, contains probability mass leading to verified correction. Forced continuation is a separate secondary intervention, not the definition of recoverability.

J-space is a candidate. J-space is tested as one distributed explanatory level. It receives no privilege unless it adds predictive or causal value beyond simpler residual-stream and confidence features.

Compute comes last. Replication, mask integrity, time-zero mapping, recoverability instrumentation, and external methodological attack come before substantive constrained adaptation. The proposal asks for the cheapest experiment capable of falsifying the programme.

Contents

How to read this paper	3
What this proposal tests	3
The Proposal in Two Minutes	4
1 Research Objective	9
2 Scope and Non-Claims	9
3 Background and Motivation	9
3.1 H-Neurons	10
3.2 Localization is not mechanism identity	10
3.3 J-space and distributed workspace-like representations	10
3.4 Hidden correctness and continuation signals are prior art	10
3.5 KV-cache as a communication medium	11
4 Operational Definitions	11
4.1 Target phenotype	11
4.2 Candidate locus	12
4.3 Selective causal contribution	12
4.4 Persistent constraint	12
4.5 Functional reconstitution and time-zero backup state	12
4.6 Compensation	13
4.7 Recoverability	13
4.8 Rescue probability	13
5 Research Questions and Hypotheses	14
5.1 RQ1 - Functional reconstitution	14
5.2 RQ2 - Recoverability prediction	14
5.3 RQ3 - Causal recoverability	14
5.4 RQ4 - Cross-level coupling	14
6 Experiment A — H-Neuron Replication and Gate 0	15
6.1 Model	15
6.2 Replication target	15
6.3 Identification - pilot lock	15
6.4 Controls - pilot lock	16
6.5 Gate 0 - pilot go/no-go rule	16
7 Experiment B — Persistent Constraint and Adaptation	17
7.1 Pilot B0 - persistent-mask integrity	17
7.2 Time-zero map is part of the pilot	17
7.3 Adaptation regimes - design retained, hyperparameters not yet frozen	17
7.3.1 B1 - Generic continuation	17
7.3.2 B2 - Targeted task adaptation	17
7.3.3 B3 - Parameter-efficient smoke test	17
7.4 Full adaptation freeze condition	18
8 Experiment C — Conditional Backup Audit and Reconstitution	18
8.1 Immutable time-zero map	18
8.2 Conditional backup discovery	18

8.3	Residual-stream information diagnostic	19
8.4	Training checkpoints	19
8.5	Evidence for adaptive change	19
9	Reconstitution Evidence Matrix	19
10	Experiment D — Pre-Commitment Recoverability	20
10.1	Pilot task families	20
10.2	Natural trajectory collection	21
10.3	Regime I0 — natural-support continuation	21
10.4	Regime I0b — minimum-one-token continuation control	21
10.5	Regime I1 — forced-compute continuation	21
10.6	Regime I2 — procedural re-check	21
10.7	Continuation count and precision	22
10.8	Prediction endpoint and split policy	22
10.9	Baselines and difficulty controls	22
10.10	J-space pilot criterion	22
11	Experiment E — Causal Recoverability	22
12	Experiment F — $H \leftrightarrow J$ Cross-Level Coupling	23
12.1	F1 - $H \rightarrow J$	23
12.2	F2 - $J \rightarrow H$	24
12.3	F3 - After adaptation	24
13	Secondary Model Organism — Proposition Substitution	24
14	Experiment G — Cross-Model State-Transfer Surplus	25
14.1	Question	25
14.2	Evaluation conditions	25
14.3	Fuser-training boundary	25
14.4	Failure interpretation	25
15	Statistical Analysis Plan	26
15.1	Freeze levels	26
15.2	Items retained as pilot-locked	26
15.3	Items reopened by adversarial review	26
15.4	Sparse-probe stability	26
15.5	Conditional-backup equivalence	27
15.6	Recoverability precision analysis	27
15.7	Held-out causal intervention	27
15.8	Bootstrap and seed reporting	27
15.9	No silent threshold repair	27
16	Compute Plan	27
16.1	Stage P1 - H-Neuron replication	27
16.2	Stage P2 - Mask/time-zero map	27
16.3	Stage P3 - Recoverability pilot	28
16.4	Stage P4 - Adaptation benchmark	28
16.5	Main allocation request	28
17	Reproducibility Package	28
18	Threats to Validity	28

18.1 Polysemanticity	28
18.2 Superposition and distributed implementation	28
18.3 Probe artifacts	29
18.4 Intervention-induced regime shift	29
18.5 Adaptation-path restriction	29
18.6 Stop-suppression artifact	29
18.7 Benchmark leakage	29
18.8 Conditional self-repair and dormant backups	29
18.9 Recoverability regime dependence	29
18.10 State-transfer fuser confounding	29
18.11 Cross-domain instability	29
19 Falsification and Stop Rules	30
Claims and Boundaries	30
20 Ethics and Interpretation Boundary	30
21 External Review Status and Remaining Questions	31
22 Current Status	31
23 Relationship to Prior Hedegreen Research Work	32
24 Planned Outputs	32
25 Conclusion	32
Data availability	33
Code availability	33
Use of generative AI tools	33
A Pilot Protocol	34
A.1 P1 — H-Neuron replication	34
A.2 P1b — Causal phenotype	34
A.3 P2 — Persistent mask / conditional t0 backup audit	34
A.4 P3 — Recoverability	34
A.5 Stop rules	35
B Freeze Register	35
C Decision Ledger	36
D Implementation Contract	37
D.1 Repository blueprint	37
D.2 Llama mask contract	38
D.3 Model contract checks	38
D.4 Source adapter rule	39
D.5 Run commands	39
D.6 No implicit defaults for scientific parameters	39
D.7 P1 outputs	39
D.8 P2 outputs	40
D.9 P3 natural trajectory representation	40

D.10 P3 continuation representation	40
D.11 P3 continuation regimes	40
D.12 Synthetic generator contract	41
D.13 Dataset fingerprint	41
D.14 Gate 0 evaluator	41
D.15 Integrity checks for pre-stop recomputation	42
D.16 Hidden-state capture	42
D.17 J-lens adapter	42
D.18 Concurrency	42
D.19 Error handling	42
D.20 Security / privacy	43
D.21 Acceptance criterion for implementation v0.1	43
References	44

1 Research Objective

The central methodological question is:

When a behaviour can be localized inside a language model, what exactly has been localized, and what happens to the function when that location is made persistently unavailable?

The proposal separates four claims that should not be silently collapsed:

1. **Predictive localization** — an internal variable predicts a behaviour.
2. **Causal contribution** — intervening on that variable changes the behaviour.
3. **Functional necessity** — the behaviour depends selectively on that variable under the tested conditions.
4. **Implementation identity** — the variable is the mechanism, rather than one implementation, readout, bottleneck, or correlated component of a larger computation.

The experiments are designed to move from weaker to stronger claims only when each preceding gate survives.

A second question concerns internal state before output commitment:

Does pre-commitment hidden state predict whether additional computation can rescue the current trajectory, beyond ordinary confidence signals?

This proposal calls that property **recoverability**.

The final programme-level question is whether the two levels are related:

Are sparse local mechanisms associated with answer commitment causally coupled to a distributed unresolved-state representation?

2 Scope and Non-Claims

This working paper proposes experiments. It does not report that any of the following have been demonstrated:

- that H-Neurons are a universal hallucination circuit;
- that hallucination is structurally necessary to next-token prediction;
- that a function will reconstitute after intervention;
- that J-space is the unique or correct representation of recoverability;
- that pre-commitment recoverability is equivalent to “knowing one is wrong”;
- that any proposed state is evidence of consciousness;
- that neuron-level localization is sufficient to identify a complete mechanism.

The earlier Hedegreen Research proposal *Compliance Is All You Learned* (Hedegreen 2026a) used iterative inference-time replacement while model weights remained fixed. That design remains useful as provenance and as a source of controls, but it is **superseded for the reconstitution claim** because a fixed-weight model cannot adaptively rebuild a mechanism.

3 Background and Motivation

3.1 H-Neurons

Gao et al. (2025) report a sparse subset of MLP intermediate neurons whose contribution profiles predict hallucination across tested model families and tasks. Their experiments also connect H-Neuron interventions to several forms of over-compliance, including invalid-premise continuation, misleading-context acceptance, sycophancy, and harmful-instruction compliance. The earlier Hedegreen Research proposal adopted the same Llama-3.1-8B-Instruct target to preserve comparability with that pipeline.

The original proposal also retained several useful methodological components:

- replication before novel intervention;
- CETT-based contribution profiling;
- threshold sensitivity;
- held-out prediction;
- sham-hook controls;
- layer/location-matched non-H controls;
- random ablation;
- fluency/perplexity monitoring;
- fixed prompt sets and random seeds;
- explicit outcome definitions.

Those components remain useful. The interpretation changes.

3.2 Localization is not mechanism identity

The broader model-editing literature warns against equating localization with mechanism identity. Hase et al. (2023) showed that causal localization did not reliably identify the most effective factual-editing locations. Later work strengthens the warning that successful behavioural editing at an internal location is not itself sufficient evidence that the behaviour is encoded there (Z. Wang and Veitch 2025).

A second problem is **conditional self-repair**. A component can appear unimportant in the intact model because a dormant backup only becomes necessary after a primary route is removed. Gong et al. (2026) formalize this as conditional circuit completion: the relevant causal quantity is how much a remaining component’s ablation effect grows once the primary set is absent. This is directly relevant to the present redundancy-versus-reconstitution problem.

The present proposal therefore requires both **selectivity** and **conditional training-time dynamics**. Ordinary marginal importance in the intact model is insufficient.

3.3 J-space and distributed workspace-like representations

Gurnee et al. (2026) introduce the Jacobian lens (J-lens), which maps earlier residual-stream activations toward final-layer representational coordinates and decodes them using the model’s output vocabulary. Their J-space construction uses sparse nonnegative combinations of an overcomplete set of J-lens vectors and is proposed as a functionally privileged representational component with several workspace-like properties.

This proposal treats J-space as a candidate explanatory level. It must compete against simpler baselines.

3.4 Hidden correctness and continuation signals are prior art

Recent work already reports that hidden states can contain information predictive of future correctness before answer generation (Cencerrado et al. 2025; Luo et al. 2026; Di Cicco 2026). The present proposal therefore does **not** claim novelty for internal correctness prediction.

Continuation itself can reveal information that forced extraction misses. H. Wang and Zhu (2026) report a detection–extraction gap in reasoning models: free continuations from partial prefixes can recover

correct answers in cases where forced extraction fails. That result strengthens the need to distinguish natural continuation from an intervention that forces additional computation.

The narrower target here is whether pre-commitment state predicts **rescue under a precisely defined continuation regime**, and whether the internal signal adds information beyond prompt difficulty, confidence and generic residual-stream features.

3.5 KV-cache as a communication medium

Fu et al. (2026) introduce Cache-to-Cache (C2C), a direct communication method in which a learned module projects and fuses one model’s KV-cache into another model’s cache. Their motivating oracle experiment is especially relevant here: they prefill a model on exemplars plus a question, discard the exemplar-cache segment, retain only the question-aligned cache, and still observe higher accuracy than direct question-only decoding at the same cache length. This supports a limited claim that the internal cache representation of a visible token sequence can carry task-relevant semantic structure that is not exhausted by token count alone.

C2C then trains a cache fuser while freezing the communicating language models. Across several model families and sizes, the authors report gains over receiver-only and text-to-text collaboration, while also showing that poor source states or poorly chosen layers can hurt performance. The result therefore does **not** imply that all internal state is useful, that KV-cache is a global workspace, or that a failed model “knows” the answer. It establishes a technically concrete route by which one model’s internal state can alter another model’s behaviour without first being compressed into intermediate text.

For the present programme, C2C motivates a downstream test of **representational surplus**: whether a failed source trajectory contains receiver-useful information in its internal cache beyond what a receiver obtains from the same visible text. Because the C2C fuser itself is learned, any such experiment must separate source-state information from information injected by the fuser’s training distribution. It is therefore an optional later branch, not evidence for the core recoverability construct.

4 Operational Definitions

4.1 Target phenotype

A **target phenotype** is a behaviour measured by a pre-specified evaluation suite.

For the pilot, the primary causal phenotype is **FaithEval — Counterfactual Context compliance rate**.

Rationale:

- it directly tests over-compliance with misleading context;
- Gao et al. use greedy decoding with a maximum of 256 new tokens;
- the primary outcome can be extracted with a rule-based parser against dataset labels, avoiding an external LLM judge for the main endpoint.

Pilot secondary phenotypes:

- FalseQA invalid-premise compliance;
- Sycophancy answer-flip rate.

Jailbreak is not a primary pilot phenotype. It remains a later replication/robustness target because its attack-success endpoint introduces an additional safety-evaluator layer and alignment-specific confounds.

The H-Neuron **identification** stage remains distinct from the behavioural phenotype stage:

- identification source: TriviaQA faithful vs hallucinated responses;
- OOD diagnostic sets: NQ-Open and BioASQ;
- NonExist retained as a diagnostic rather than a pass/fail gate for Llama-3.1-8B because the source paper reports weak performance for that model on this particular set.

4.2 Candidate locus

A **candidate locus** is an internal neuron, neuron population, direction, circuit, or representational component selected by a pre-specified localization procedure.

A candidate locus is not assumed to be the complete mechanism.

4.3 Selective causal contribution

A locus demonstrates **selective causal contribution** when intervention on that locus changes the target phenotype more strongly than matched control interventions while preserving sufficient general capability to rule out gross model damage as the explanation.

4.4 Persistent constraint

A **persistent constraint** is an intervention kept active on every relevant forward pass during adaptation.

For the first H-Neuron experiment, the candidate implementation is a fixed mask on selected MLP intermediate dimensions applied after the gated activation and before the down-projection.

The implementation must be audited to confirm that the constrained dimensions remain unavailable throughout training.

4.5 Functional reconstitution and time-zero backup state

R0 establishes **selective causal contribution**, not mechanism identity. Stronger claims require additional gates and a conditional time-zero backup audit.

The core R0–R4 sequence remains:

- **R0 — Selective initial effect:** the original target has a selective causal effect beyond matched controls.
- **R1 — Phenotype loss:** persistent constraint measurably reduces the target phenotype before adaptation.
- **R2 — Training-dependent recovery:** the phenotype recovers during adaptation relative to masked no-training and other relevant controls.
- **R3 — Training-dependent replacement change:** a candidate replacement becomes more predictive and/or more conditionally causal across training checkpoints than it was at time zero.
- **R4 — Selective causal dependence:** intervention on the candidate replacement selectively damages the recovered phenotype beyond matched control interventions.

R0–R4 are interpreted together with a separate **B0 conditional backup state** measured immediately after the original locus is constrained and before adaptation:

- **B0-A — Near-null backup:** no candidate alternative carries more than a pre-registered practically negligible conditional causal effect.
- **B0-B — Weak backup:** a candidate route has non-negligible but limited conditional causal effect at time zero.
- **B0-C — Strong pre-existing backup:** an alternative route already carries substantial conditional causal effect before adaptation.

The strongest label, **de novo / near-null functional reconstitution**, is reserved for B0-A plus R0–R4 and cross-seed replication. B0-B followed by a large training-dependent increase is classified as **adaptive amplification of pre-existing redundancy**. B0-C is classified as **self-repair / pre-existing backup**, not reconstitution.

Version 0.6 freezes the *rule* for the B0-A practical-equivalence margin before adaptation. Let Δ_{R1} be the absolute pre-adaptation phenotype loss caused by the persistent H-mask. Define

$$\delta_{B0} = \min(0.02, 0.20 |\Delta_{R1}|),$$

where effects are expressed as absolute proportions. Because Gate 0 already requires at least a five-percentage-point phenotype shift, this rule yields a candidate equivalence margin between one and two percentage points. B0-A is available only when a paired 90% confidence interval for the candidate backup’s conditional causal risk difference lies wholly inside $[-\delta_{B0}, +\delta_{B0}]$. This is the paired-binary analogue of an equivalence test; a generic two-sample *t*-test is not used for binary compliance outcomes.

The strongest label is also **power-gated**: if the fixed evaluation set cannot provide at least 80% power for the frozen equivalence margin under the pre-adaptation paired design, B0-A remains unresolved and de novo / near-null reconstitution cannot be claimed. The evaluation set must then be enlarged or the strongest label remains unavailable. The margin rule may not be changed after any adaptation result is observed.

4.6 Compensation

Compensation is behavioural recovery through an adaptation mechanism whose allowed parameter space or architecture materially predetermines where replacement may occur, without sufficient evidence for broader network reorganization.

A LoRA-only recovery may therefore qualify as compensation rather than full reconstitution unless the interpretation is explicitly restricted to the permitted adaptation subspace.

4.7 Recoverability

Version 0.5 separates three intervention-relative constructs rather than treating hard stop suppression as the definition of recoverability.

Natural-support recoverability asks whether, from the exact pre-stop prefix and under a frozen stochastic decoding policy with termination still allowed, the model’s own continuation distribution assigns non-zero probability to a verified correction.

Forced-compute recoverability asks whether a verified correction can be obtained when termination is constrained and additional bounded computation is forced. This is a distinct intervention and is not interpreted as the model’s natural continuation tendency.

Procedural re-check recoverability asks whether a fixed instruction to continue checking, without new task facts, produces a correction. This is a third regime and is reported separately.

Recoverability is therefore decoding-policy and intervention-relative. It is not defined as an intrinsic mental state or as evidence that the model “knows” it is wrong.

4.8 Rescue probability

For a pre-commitment prefix state s , continuation regime I , decoding policy D , token budget B , and K sampled continuations:

$$\hat{p}_{rescue}(s; I, D, B) = \frac{1}{K} \sum_{k=1}^K \mathbb{I}[\text{continuation}_k \text{ reaches verified correction}]$$

The continuation count K remains open in v0.6 until the pre-specified precision/cost simulation is frozen. A pre-compute precision/cost simulation will compare $K \in \{8, 16, 32\}$ before P3 result collection. The token budget remains a pilot design parameter and is reported separately for each continuation regime.

5 Research Questions and Hypotheses

5.1 RQ1 - Functional reconstitution

If a selectively causal H-Neuron locus is made persistently unavailable, how does the target phenotype recover, if at all, and what conditional backup state existed before learning?

The evidence is reported on orthogonal axes rather than forced into one mutually exclusive label. Derived labels include:

Elimination: target behaviour remains reduced while general capability recovers.

Self-repair / pre-existing backup: a strong conditional backup is already causal at B0-C before adaptation.

Adaptive amplification: a weak B0-B route becomes substantially more predictive and causally important during training.

Distributed redistribution: behaviour recovers without a compact selectively causal replacement.

Restricted-subspace compensation: recovery occurs within a deliberately restricted adaptation mechanism such as PEFT/LoRA.

De novo / near-null functional reconstitution: B0-A plus R0–R4 and cross-seed replication.

Degeneration: useful capability cannot recover under the constraint.

No one outcome is designated as the desired result.

5.2 RQ2 - Recoverability prediction

Does pre-commitment hidden state predict rescue probability beyond simple output-confidence and task-difficulty baselines?

H2a: simple confidence features explain most recoverability variance.

H2b: generic residual-stream probes add predictive value.

H2c: J-space-derived features add predictive value beyond both.

5.3 RQ3 - Causal recoverability

Can intervention on a recoverability-associated internal representation selectively alter useful continuation or correction, rather than merely increasing verbosity or hesitation?

5.4 RQ4 - Cross-level coupling

Are H-Neuron-associated commitment mechanisms and a distributed unresolved-state representation causally coupled?

6 Experiment A — H-Neuron Replication and Gate 0

6.1 Model

Primary model:

meta-llama/Llama-3.1-8B-Instruct

Rationale:

- continuity with the earlier Hedegreen proposal;
- continuity with the published H-Neuron pipeline;
- tractable open-weight model size.

The exact model revision must be pinned before execution.

6.2 Replication target

The first experiment is not novel.

It must establish that the source signal can be reproduced sufficiently to justify later causal work.

Required outputs:

- number and layer distribution of identified H-Neurons;
- held-out hallucination/over-compliance prediction performance;
- sensitivity to identification threshold;
- intervention effect on target phenotype;
- general language/perplexity impact;
- variance across seeds.

6.3 Identification - pilot lock

The pilot follows the source-paper structure:

- train sparse L1 logistic regression on CETT contribution profiles from TriviaQA;
- compute contributions on answer-token positions;
- compare against the same number of randomly selected neurons;
- evaluate transfer to NQ-Open and BioASQ;
- retain NonExist as a diagnostic.

Coefficient-threshold sensitivity is frozen for the pilot at:

{0.0005, 0.001, 0.002}

Primary threshold: **0.001**.

Pilot sample rule:

- deterministic sample of 400 eligible examples per identification/evaluation set when at least 400 are available;
- otherwise use all eligible examples;
- sampling seed = 42;
- two replication generation seeds, 314 and 2718, in addition to seed 42 where applicable.

The official H-Neuron paper reports, for Llama-3.1-8B, H-vs-random accuracy differences of +14.0 percentage points on TriviaQA, +10.3 on NQ-Open, and +13.1 on BioASQ. A conservative independent-proportions approximation requires roughly 359 examples per group for 80% power at $\alpha = 0.05$ to detect

the smallest of these reported differences; the pilot therefore targets 400 eligible evaluation examples where available. Because H and random classifiers are evaluated on the same examples, final uncertainty reporting uses paired resampling / paired tests rather than treating predictions as independent.

The sparse-classifier implementation is pinned to the official source pipeline and commit. The source paper uses L1-regularized logistic regression and chooses regularization strength by grid search over predictive accuracy plus intervention-side model performance. Version 0.6 therefore does not invent a new fixed solver/ C pair. Instead, the exact upstream grid, solver, feature preprocessing, tolerance, and scikit-learn version are recorded and frozen from the pinned implementation before result-bearing P1. At least ten resampled/refit identification runs are used to report neuron-selection frequency and pairwise Jaccard overlap. No arbitrary Jaccard threshold defines success; if unit identities are unstable but population-level causal effects replicate, the target is downgraded from individual-neuron identity to a population/subspace description.

6.4 Controls - pilot lock

Minimum controls:

1. sham hook;
2. random neuron mask with matched neuron count;
3. same-layer non-H-neuron mask;
4. contribution-matched non-H-neuron mask;
5. H-Neuron soft suppression sensitivity ($\alpha=0.5$);
6. H-Neuron hard suppression ($\alpha=0.0$);
7. WikiText-103 or an equivalent fixed held-out LM corpus for perplexity/general-language damage.

All control conditions use identical instrumentation.

6.5 Gate 0 - pilot go/no-go rule

The pilot proceeds to persistent-mask/time-zero mapping only if all of the following are true:

G0-A — Identification advantage

On TriviaQA, the H-Neuron classifier must exceed the matched random-neuron classifier by at least **8 percentage points** in held-out accuracy, and a paired 95% confidence interval for the difference must exclude zero. The direction of advantage must also be positive on NQ-Open and BioASQ.

The 8-point value remains a *compute-screening threshold*, not the scientific definition of replication. The source Llama-3.1-8B table reports a +14.0-point TriviaQA advantage, with +10.3 on NQ-Open and +13.1 on BioASQ. The paper reports the full source-vs-replication effect sizes and intervals rather than treating the 8-point gate as evidence by itself.

G0-B — Primary causal phenotype

On FaithEval Counterfactual Context, hard H-Neuron suppression must reduce compliance rate by at least **5 percentage points** relative to the unmodified baseline.

G0-C — Selectivity against matched controls

The FaithEval compliance reduction under H-Neuron suppression must exceed the contribution-matched non-H-neuron reduction by at least **3 percentage points**.

A stratified bootstrap 95% interval for the H-vs-matched-control difference must exclude zero.

G0-D — Damage boundary

Relative perplexity degradation on the fixed LM corpus must remain below **10%** for the pilot to be interpreted as selective enough to justify the next stage.

If behavioural benefit appears only with >10% perplexity damage, classify the pilot as **AMBIGUOUS / DAMAGE-CONFOUNDED** rather than passing Gate 0.

These are pilot gates, not final-paper inferential thresholds. They may be revised exactly once after external review and before any adaptation result is observed.

7 Experiment B — Persistent Constraint and Adaptation

7.1 Pilot B0 - persistent-mask integrity

v0.2 freezes a **no-training mask pilot** before any adaptation run.

For each selected H-Neuron dimension j , define $m_j=0$; all other dimensions use $m_j=1$.

At the MLP intermediate activation: $h' = m * h$.

The mask is applied after the gated intermediate activation and before the down-projection.

Before training, verify on a fixed audit batch that:

- every constrained dimension is exactly zero at the masked point;
- the mask is active at every intended layer and token position;
- forward-hook / compiled-model paths do not bypass it;
- control masks use identical code;
- gradients through the masked activation do not restore a non-zero contribution at that point.

A unit test must fail if any masked dimension exceeds an absolute activation tolerance of **1e-7** at the audited tensor.

7.2 Time-zero map is part of the pilot

Immediately after masking and before adaptation:

- rerun localization;
- rank the strongest alternative loci;
- measure their predictive signal;
- intervene on the top alternatives;
- save the full candidate-locus table.

This t0 table becomes immutable provenance for the later redundancy test.

7.3 Adaptation regimes - design retained, hyperparameters not yet frozen

7.3.1 B1 - Generic continuation

Continue next-token training on a controlled general-text corpus.

7.3.2 B2 - Targeted task adaptation

Continue training on a held-out task distribution exposing the target phenotype.

7.3.3 B3 - Parameter-efficient smoke test

LoRA/PEFT may be used only to test plumbing, logging, checkpointing and whether any recovery signal is measurable cheaply.

A LoRA-only recovery is labelled **COMPENSATION-IN-PERMITTED-SUBSPACE** unless a later broad-adaptation experiment supports a stronger claim.

7.4 Full adaptation freeze condition

Optimizer, learning rate, token budget, checkpoint schedule and parameter-freezing policy are **not** frozen in v0.2.

They are frozen only after:

1. Gate 0 passes;
2. the B0 mask audit passes;
3. a short throughput/memory benchmark is recorded;
4. at least one external technical review addresses adaptation breadth.

No main-run adaptation result may be collected before that freeze.

8 Experiment C — Conditional Backup Audit and Reconstitution

8.1 Immutable time-zero map

The t0 map created under the persistent H-mask is the reference against which all later replacement claims are judged.

For every candidate alternative route recorded at t0, retain:

- layer / neuron / direction / structured-group identifier;
- predictive score;
- marginal causal intervention effect;
- conditional causal effect given removal of the original target;
- activation/contribution statistics;
- rank among matched candidates.

8.2 Conditional backup discovery

Ordinary intact-model importance is insufficient when self-repair is possible. Following the conditional logic developed by Gong et al. (2026), P2 therefore asks which remaining components become more causally important *because* the primary H-associated set is absent.

Where exact neuron-scale conditional co-ablation is computationally feasible, it is the preferred backup-discovery analysis. Where it is prohibitive, the approximation must be explicit: first use structured groups or a bounded candidate pool selected by pre-specified localization/contribution criteria, then perform conditional co-ablation and causal validation inside that pool.

Version 0.6 adds two validation requirements before a bounded-pool B0-A result can support the strongest claim. First, the local CoAx adapter must reproduce the published GPT-2-small IOI backup-recovery benchmark closely enough to demonstrate that the implementation can recover a known self-repair structure. Second, the Llama backup audit must include a **candidate-pool expansion sensitivity analysis**: enlarge the candidate pool across pre-specified sizes / structured groups and test whether the B0 conclusion stabilizes. A reference benchmark validates the implementation, but it does not prove that the bounded pool is complete in the target LLM. If the B0 classification changes materially as the pool expands, B0-A is unresolved and the de novo label is unavailable.

The purpose is not to import CoAx as a universal answer, but to measure the specific confound the original design lacked: dormant routes that become necessary only after the primary route is removed.

8.3 Residual-stream information diagnostic

At baseline, t0 and adaptation checkpoints, fit held-out residual-stream probes for the target phenotype. This is a diagnostic, not a pass/fail gate.

High residual decodability after local masking can mean that information remains available even though the removed local route was causally important. It does not by itself show that an alternative route can use that information. The key distinction is therefore:

- information retained, route lost;
- information retained, alternate route already conditionally causal;
- information or causal routing redistributed during adaptation.

No arbitrary residual-probe AUROC threshold is used to define success or failure. High decodability can show that information remains present; it cannot by itself show that an alternate route already uses that information causally.

At every adaptation checkpoint, report both (a) the frozen original H-Neuron detector applied without refitting and (b) a checkpoint-refit detector trained under the same pinned pipeline on the checkpoint model's activations. This separates loss of validity of the original detector from genuine emergence of a new predictive population. Large detector-set changes are reported as representation drift and must be resolved by causal tests rather than interpreted as reconstitution by themselves.

8.4 Training checkpoints

The main-run checkpoint schedule is not yet numeric. The schedule rule remains:

- t0 before adaptation;
- at least four intermediate checkpoints;
- final checkpoint;
- approximately logarithmic spacing in consumed training tokens.

Exact token counts are frozen after the throughput pilot and before main-run adaptation.

8.5 Evidence for adaptive change

A replacement route is not classified as new merely because it is detectable after training. R3 requires change relative to the candidate's own t0 conditional baseline.

Pre-specified evidence axes include:

- held-out predictive AUROC / accuracy;
- marginal causal effect;
- conditional causal effect under the original mask;
- rank among matched candidates;
- residual-stream representation change;
- stability across seeds.

Mediation-style analyses remain exploratory unless frozen before the main run.

9 Reconstitution Evidence Matrix

The primary report uses orthogonal evidence axes. Derived labels summarize the pattern but do not replace the underlying measurements.

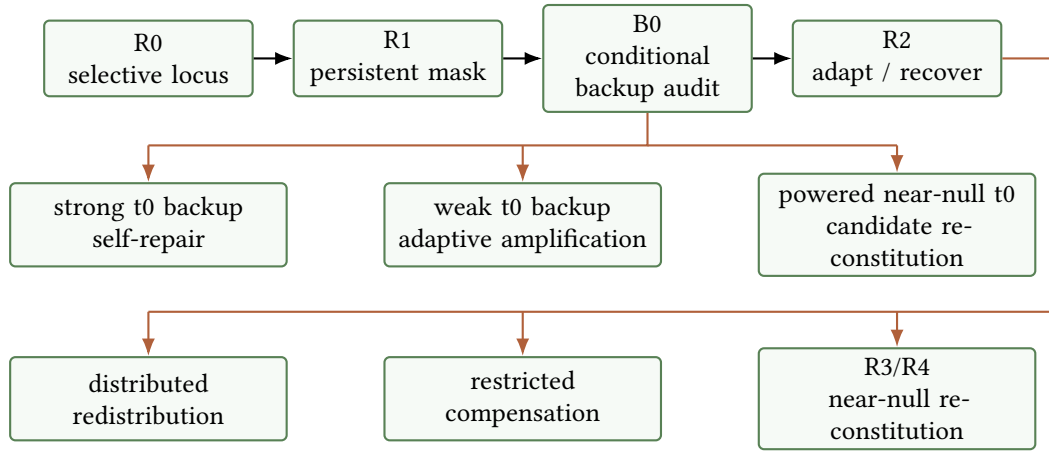


Figure 1: Version 0.6 makes the conditional time-zero backup audit a branch point. The strongest near-null reconstitution path is available only after a powered practical-equivalence test and later training-dependent R3/R4 causal evidence.

Axis	States	Interpretation
Phenotype recovery	none / partial / strong	Did the target behaviour return with learning?
B0 backup strength	near-null / weak / strong	Was a conditionally causal backup already available before adaptation?
Representation change	local / distributed / none detected	Where did predictive structure change?
Conditional causal replacement	absent / weak / strong	Did a later route become necessary specifically under the original constraint?
Adaptation restriction	broad / PEFT-limited	Was the replacement search space artificially constrained?
General-capability damage	low / moderate / high	Could apparent effects be explained by model degradation?

Derived labels:

- **Self-repair / pre-existing backup:** strong B0 conditional causal effect before adaptation.
- **Adaptive amplification:** weak B0 backup with substantial training-dependent causal growth.
- **Distributed redistribution:** phenotype recovery without a compact selectively causal replacement.
- **Restricted-subspace compensation:** recovery confined to a pre-specified adaptation subspace.
- **De novo / near-null functional reconstitution:** B0 near-null plus R0–R4 and cross-seed replication.
- **Elimination / degeneration / ambiguous:** reported when supported by the corresponding axes.

Ambiguous outcomes remain ambiguous.

10 Experiment D — Pre-Commitment Recoverability

10.1 Pilot task families

The recoverability pilot begins with deterministic synthetic task families whose answers are computed by code rather than an LLM judge:

1. integer arithmetic with controlled operator depth;
2. modular arithmetic;
3. short symbolic / Boolean logic problems.

Difficulty is calibrated on a development generator before probe fitting. The primary analysis includes difficulty covariates and reports performance across difficulty strata; a signal confined to a narrow calibrated band is described as such.

A fixed public reasoning benchmark with exact-answer scoring may be added as secondary external validity after the synthetic pipeline is frozen.

10.2 Natural trajectory collection

For each task, collect one deterministic natural trajectory under the frozen natural decoding policy. Save token IDs, logits, hidden states, task difficulty parameters, and the exact prefix immediately before natural termination/finalization.

10.3 Regime I0 — natural-support continuation

I0 is the primary recoverability construct.

For each naturally incorrect or incomplete trajectory:

1. restore the exact pre-stop prefix;
2. leave EOS / normal termination available;
3. sample K independent continuations under a frozen stochastic decoding policy;
4. provide no new factual task information;
5. score the last valid final answer with the deterministic task checker.

If EOS is selected immediately, that continuation is a non-rescue. I0 therefore estimates probability mass assigned by the model’s own continuation distribution to a verified correction *under its ordinary termination policy*. It is not interpreted as the total amount of computation that could in principle be recovered if the termination barrier were altered.

10.4 Regime I0b — minimum-one-token continuation control

I0b isolates the termination-policy confound. From the same pre-stop prefix, EOS is disallowed for exactly the first generated token and then immediately re-enabled. All other decoding parameters match I0.

The contrast I0b–I0 estimates how much rescue probability changes when the model is required to cross only the immediate stop barrier. A large difference indicates that natural-support recoverability is strongly entangled with termination policy and must be reported as such. I0b is a control, not a replacement for I0.

10.5 Regime I1 — forced-compute continuation

I1 retains hard stop suppression as a **secondary** intervention. EOS/stop is made unavailable for a bounded window and the model is forced to generate additional tokens. This measures what can be recovered under forced compute, not natural-support recoverability.

10.6 Regime I2 — procedural re-check

I2 adds a fixed procedural instruction containing no new task facts:

“Continue checking your reasoning. You may revise the final answer.”

I0, I0b, I1 and I2 are not pooled into one rescue label. Their agreement or disagreement is itself a result

about termination policy and intervention dependence. The free-continuation procedure described in the detection–extraction literature is represented here by I0; it is an outcome-generating regime, not a predictor baseline that a hidden-state probe can simply “beat.”

10.7 Continuation count and precision

The continuation count remains reopened in v0.6. Before result-bearing P3 collection, simulate $K \in \{8, 16, 32\}$ under plausible rescue rates and the planned number of incorrect trajectories. Select the **smallest** K for which (a) the Monte Carlo standard error of the planned held-out mean negative log likelihood is at most 0.01 and (b) doubling K changes that estimated standard error by less than 20%. If no candidate satisfies both criteria, increase K beyond 32 or reduce the precision claim before data collection. The selected value and simulation code are frozen before P3. This population-level criterion is used instead of requiring a narrow confidence interval for every individual trajectory, which would be prohibitively expensive and unnecessary for the hierarchical primary analysis.

10.8 Prediction endpoint and split policy

The primary predictive comparison uses continuation-level Bernoulli outcomes or an equivalent hierarchical/binomial model with examples treated as the repeated-measure unit. Held-out negative log likelihood / deviance is the primary scoring rule.

Synthetic tasks are split by generator seed and template family. Discovery, validation and final test data remain separate; no probe hyperparameter is selected on the final test split.

10.9 Baselines and difficulty controls

Before J-space, compare against:

1. output entropy;
2. top-1/top-2 logit margin;
3. EOS/stop-token probability;
4. response length;
5. synthetic task difficulty parameters;
6. prompt-only lexical/length features;
7. question-only hidden-state probe before answer generation;
8. generic residual-stream linear probe at the pre-stop state.

Report rescue prediction as a function of task difficulty and include interaction terms or stratified curves sufficient to show whether the signal generalizes beyond the calibrated difficulty band.

10.10 J-space pilot criterion

J-space remains optional. It is only interpreted as adding explanatory value if it improves held-out prediction beyond the best simpler baseline with a pre-specified effect criterion and consistent direction across task families. The previous 2% screening rule remains reopened in v0.6 and must be re-frozen after the revised baseline stack is benchmarked.

11 Experiment E — Causal Recoverability

Proceed only if Experiment D finds a robust predictive state beyond simpler baselines. For confirmatory progression, the held-out log-loss improvement over the best non-target baseline must have a bootstrap 95% confidence interval excluding zero overall, the direction of improvement must be consistent across all three synthetic task families, and the effect must replicate on the secondary external-validity set

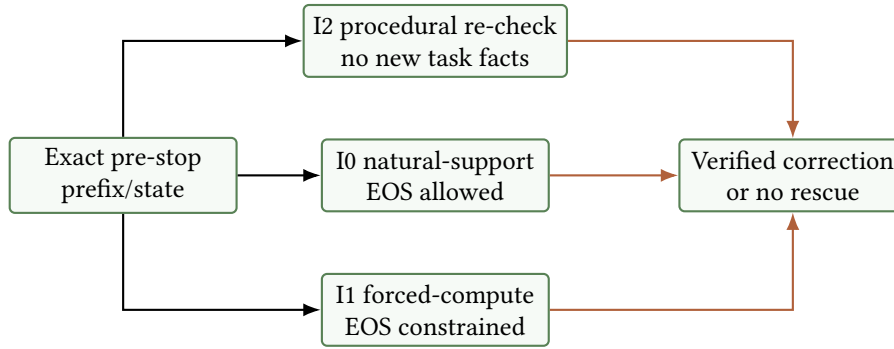


Figure 2: Recoverability is regime-dependent. Natural-support continuation is the primary construct; forced-compute and procedural re-check continuations are separate interventions and are not pooled into one label.

before a broad recoverability-state claim is made. No arbitrary 5% improvement threshold is imposed solely because an adversarial reviewer proposed it.

To avoid circularity, representation discovery and causal testing are separated:

1. fit the recoverability representation/probe on the discovery split;
2. choose intervention direction and strength on validation data;
3. measure causal effects on untouched test data;
4. perform a cross-family test in which a direction discovered on two synthetic task families is intervened on in the held-out third family.

Candidate interventions include amplification, suppression and matched-state patching.

Primary outcomes include correction probability, final accuracy, continuation length, unnecessary continuation on already-correct cases, global hesitation/abstention and generic language degradation.

A successful causal result should selectively improve recoverable cases rather than merely increase thinking length or caution.

12 Experiment F — $H \leftrightarrow J$ Cross-Level Coupling

Proceed only after the H-Neuron and recoverability/J-space branches independently survive their own gates.

A raw $H \rightarrow J$ effect is not sufficient because H-associated MLP outputs feed the residual stream that the J-lens reads. The bridge experiment therefore asks for **selective coupling beyond generic downstream propagation**.

12.1 F1 - $H \rightarrow J$

Compare H-Neuron intervention against:

- contribution-matched non-H neurons;
- random same-layer neurons;
- effect/norm-matched perturbations where feasible;
- an isotropic random residual-stream direction injected at the same layer and scaled to match the induced residual-norm change of the H intervention.

Measure whether J-space changes predict target behavioural change beyond generic residual disturbance. No finite control set can prove that all downstream propagation has been removed; the goal is to bound

generic propagation strongly enough that any remaining selective coupling is interpretable.

12.2 F2 - J → H

Compare the candidate J/recoverability direction against:

- random J-space directions;
- norm-matched non-J residual directions;
- shuffled recoverability directions.

Measure H-Neuron activation/contribution, over-compliance and stopping behaviour.

12.3 F3 - After adaptation

Repeat only after Experiment B/C. A high-value pattern would be a stable high-level recoverability state paired with a changed low-level implementation, but this interpretation requires the matched propagation controls above.

The bridge experiment may be dropped entirely if J-space does not add value beyond simpler baselines.

13 Secondary Model Organism — Proposition Substitution

A separate controlled task examines claim substitution.

Example structure:

- user supplies a narrow proposition;
- model is asked to analyse it;
- output is labelled as SAME / STRONGER / WEAKER / DIFFERENT / TEMPORAL-INFLATION / CERTAINTY-INFLATION / INVENTED-CLAIM.

Research question:

Before a substituted claim is generated, does hidden state contain information predictive of the mismatch?

Advantages:

- source proposition is exactly known;
- no external factual knowledge is needed;
- error is relational and labelable.

Controls:

- minimal pairs;
- paraphrase-balanced prompts;
- held-out templates;
- lexical-feature baselines;
- cost to faithful answers;
- generic-caution controls.

This experiment may become an independent paper if the signal is sufficiently clean.

14 Experiment G — Cross-Model State-Transfer Surplus

This optional downstream experiment is motivated by Cache-to-Cache communication (Fu et al. 2026). It is not required for the H-Neuron reconstitution programme and does not determine the primary recoverability result.

14.1 Question

For source trajectories that end incorrectly or incompletely:

Does a separate receiver perform better when given transformed internal KV-cache from the failed source than when given the same visible text without that cache transfer?

A positive result would support a bounded statement: the source model’s internal state contained task-relevant information that was usable by another model and not fully reproduced by the receiver from the visible text alone. It would **not** show that the source model could have used that information itself, that the source was consciously aware of it, or that the cache represents the same construct as J-space or recoverability.

14.2 Evaluation conditions

On a frozen held-out task set, evaluate at least:

1. **Receiver-only text:** receiver gets the visible source prefix / output text only.
2. **Text-to-text handoff:** source produces an explicit intermediate analysis or message which the receiver receives with the task.
3. **Cache-assisted:** receiver gets the same visible text plus transformed/fused source KV-cache.
4. **Shuffled-cache control:** cache is taken from a different task instance matched as closely as practical for length and intervention point.
5. **Wrong-source control:** source cache comes from a deliberately irrelevant or misleading source state where task design permits.

The principal contrast is cache-assisted minus receiver-only text, with text-to-text handoff as the explicit verbal-communication baseline. Shuffled and wrong-source conditions test whether gains arise from generic extra activation or task-specific transferred state.

14.3 Fuser-training boundary

The C2C paper trains the fuser while freezing the source and receiver models. That makes the fuser a potential explanatory confound in the present setting. To avoid answer leakage:

- the fuser training corpus must be disjoint from the held-out evaluation task instances;
- where synthetic tasks are used, generator seeds and task templates used for evaluation must be held out from fuser training;
- the fuser is frozen before any state-transfer evaluation label is observed;
- receiver-only, text-to-text and cache-assisted conditions use the same receiver weights;
- cache-pair alignment and layer-gating decisions are frozen before the final evaluation split.

A stronger generalization test trains the fuser on generic instruction data or on task families different from the evaluation family and then evaluates cache transfer on held-out domains.

14.4 Failure interpretation

If cache-assisted transfer does not outperform text-only transfer, the result does not show that the source internal state lacked task-relevant information. It may reflect failed cross-model alignment, destructive fusion, or an inadequate fuser. Conversely, if cache-assisted transfer improves performance, the gain

must survive shuffled-cache and wrong-source controls before being interpreted as source-state-specific semantic transfer.

The strongest permissible claim is therefore about **receiver-usable representational surplus**, not latent knowledge in the source model.

15 Statistical Analysis Plan

15.1 Freeze levels

PILOT-LOCKED items may not change during a result-bearing pilot except under a new protocol version. **MAIN-RUN-TBD** items remain intentionally unresolved until pilot benchmarking and external review. **EXPLORATORY** analyses cannot determine the primary outcome class.

15.2 Items retained as pilot-locked

Retained:

- primary H causal phenotype = FaithEval Counterfactual Context compliance;
- identification threshold sensitivity = {0.0005, 0.001, 0.002}, primary 0.001;
- H hard suppression = alpha 0.0; soft sensitivity = alpha 0.5;
- Gate 0 behavioural selectivity and damage checks, explicitly treated as compute-screening gates rather than scientific replication thresholds;
- deterministic synthetic recoverability task families and disjoint generator-seed splits;
- natural trajectory decoding temperature = 0;
- no silent threshold repair.

15.3 Items reopened by adversarial review

Version 0.6 re-freezes several items after the second adversarial review and keeps others explicitly open:

- I0 remains the primary natural-support regime; I0b is added as a one-token termination-barrier control;
- I1 remains secondary forced compute and I2 remains procedural re-check;
- the B0 equivalence-margin *rule* is frozen as $\delta_{B0} = \min(0.02, 0.20|\Delta_{R1}|)$, with power required for the strongest label;
- conditional-backup discovery now requires implementation validation plus candidate-pool expansion sensitivity;
- recoverability continuation count K remains open only until the pre-specified precision/cost simulation is run;
- J-space screening threshold remains open until the revised baseline stack is benchmarked;
- outcome reporting remains an evidence matrix with derived labels, not a forced single class.

15.4 Sparse-probe stability

Freeze the complete sparse H-Neuron probe pipeline before P1 result-bearing runs by pinning the official H-Neurons repository commit and recording its solver, grid-search procedure for regularization strength, train mode, feature construction, convergence tolerance and software version. The official source is replicated rather than replaced by an arbitrary new C value.

Refit across at least ten resampled / independently generated identification datasets where feasible and report selection frequency and pairwise Jaccard overlap. No arbitrary Jaccard pass threshold is imposed: sparse correlated features can yield unstable unit identity even when a population-level signal is stable. Confirmatory neuron-level interpretation requires that independently fit H masks produce

a reproducible causal phenotype effect; otherwise the target is downgraded to a population/subspace description.

15.5 Conditional-backup equivalence

The B0-A margin rule is frozen in v0.6 as $\delta_{B0} = \min(0.02, 0.20|\Delta_{R1}|)$. Equivalence is assessed on the paired conditional causal risk difference using a 90% confidence interval. The strongest label is unavailable unless the design has at least 80% power for this frozen margin. If power is inadequate, increase the evaluation sample before adaptation or report B0 as unresolved. This makes the near-null boundary falsifiable without pretending that an arbitrary one-percentage-point margin is estimable with a small pilot sample.

15.6 Recoverability precision analysis

Before P3, simulate the uncertainty and compute cost of $K = 8, 16, 32$ continuations over plausible rescue-rate ranges and the planned test-set size. Select the smallest K that yields Monte Carlo standard error at most 0.01 for the planned held-out mean negative log likelihood and for which doubling K changes that estimated standard error by less than 20%. Freeze the simulation, result and selected K before P3.

Primary rescue prediction is evaluated on continuation-level Bernoulli outcomes or a hierarchical/binomial equivalent, with example-level dependence accounted for. Difficulty covariates and out-of-band validation are mandatory.

15.7 Held-out causal intervention

Recoverability-direction discovery, intervention tuning and confirmatory causal testing use separate splits. A direction fit on the causal test examples cannot support a confirmatory intervention claim.

15.8 Bootstrap and seed reporting

For FaithEval compliance-rate contrasts, retain stratified bootstrap confidence intervals. Report every seed individually; do not average away failed seeds.

15.9 No silent threshold repair

If a frozen pilot gate fails, do not modify the threshold and rerun while retaining the same protocol status. A revised protocol requires a new version and ledger entry before new result-bearing runs.

16 Compute Plan

v0.2 still refuses to invent a full-training GPU budget.

The old inference-only estimate is obsolete.

16.1 Stage P1 - H-Neuron replication

Single high-memory GPU where feasible. Record wall-clock time, peak VRAM, activation-cache size and per-benchmark runtime.

16.2 Stage P2 - Mask/time-zero map

Same hardware class if feasible. No adaptation.

16.3 Stage P3 - Recoverability pilot

Inference-only generation and hidden-state capture. Estimate storage separately from GPU time because $K=8$ counterfactual continuations multiply generation volume.

16.4 Stage P4 - Adaptation benchmark

Only after Gate 0/B0. Run a deliberately short benchmark sufficient to measure tokens/sec/GPU, peak training memory, optimizer-state memory, checkpoint write time and checkpoint size. No scientific reconstitution claim is made from this benchmark.

16.5 Main allocation request

Only after P4, calculate GPU-hours as GPUs x hours/run x conditions x seeds, with experimental and control runs listed separately.

17 Reproducibility Package

The intended release package includes:

- model revision hashes;
- environment lockfile;
- experiment configs;
- prompt/task datasets where licensing permits;
- train/validation/test splits;
- random seeds;
- neuron masks;
- checkpoint metadata;
- logs;
- evaluation scripts;
- probe training code;
- analysis notebooks;
- negative results;
- machine-readable decision ledger.

The final paper should distinguish:

- source-derived code;
 - modified source code;
 - new Hedegreen Research code.
-

18 Threats to Validity

18.1 Polysemanticity

Neuron-level interventions may affect several unrelated functions.

Matched controls and general-capability measurements reduce but do not eliminate this problem.

18.2 Superposition and distributed implementation

A sparse H-Neuron signal may be only one projection of a distributed computation.

A failure to find a replacement neuron is not evidence that no replacement computation exists.

18.3 Probe artifacts

Linear probes may exploit correlates that the model does not use.

Causal follow-up is required.

18.4 Intervention-induced regime shift

Persistent masking may move the model into a distribution not represented by the original localization procedure.

At every checkpoint, the analysis therefore reports both the frozen original detector and a checkpoint-refit detector under the same pinned pipeline. A divergence between them is evidence of detector/representation drift, not automatically evidence of a new mechanism. Causal testing remains required.

18.5 Adaptation-path restriction

PEFT methods may predetermine where compensation can occur.

Claims must match the allowed parameter space.

18.6 Stop-suppression artifact

Preventing termination may create unnatural text or alter the task itself.

Recoverability must be tested with at least one alternative commitment intervention.

18.7 Benchmark leakage

Source benchmarks may overlap with model training or tuning data.

Where possible, newly generated or held-out synthetic tasks should supplement standard benchmarks.

18.8 Conditional self-repair and dormant backups

A backup component can appear causally irrelevant in the intact model and become important only after a primary component is removed. Marginal ablation therefore underestimates some redundancy. The t0 protocol now includes conditional backup discovery and causal validation rather than relying on intact-model scores alone (Gong et al. 2026).

18.9 Recoverability regime dependence

Natural-support, forced-compute and procedural re-check continuations define different counterfactual regimes. A rescue effect in one regime need not generalize to another. This is an interpretation boundary rather than noise to be averaged away (H. Wang and Zhu 2026).

18.10 State-transfer fuser confounding

A trained cache fuser can itself learn task structure. Any Experiment G gain may therefore arise from the fuser's training distribution rather than information uniquely supplied by the source cache. Disjoint fuser training, frozen evaluation, shuffled-cache controls and cross-domain transfer tests reduce this ambiguity but do not eliminate it. State-transfer results must remain conditional on the learned alignment mechanism.

18.11 Cross-domain instability

H-Neuron identity may differ across domains (S. Vaddi and P. Vaddi 2026).

The experiment should avoid treating one domain-specific population as universal.

19 Falsification and Stop Rules

Stop or narrow the programme if:

1. H-Neuron replication fails materially.
2. Matched controls reproduce the same behavioural effect.
3. Persistent masking does not reduce the target phenotype.
4. Recovery occurs without any training-dependent representational or causal change.
5. Candidate replacement loci were already equally causal at time zero.
6. J-space adds no value beyond simpler recoverability baselines.
7. Recoverability intervention merely increases verbosity or hesitation.
8. Cross-level $H \leftrightarrow J$ effects fail to replicate.
9. proposition-substitution signal reduces to surface lexical cues.

These are not publication failures.

They are valid research outcomes.

Claims and Boundaries

THE PROPOSAL TESTS	THE PROPOSAL DOES NOT ESTABLISH
Whether recovery is explained by self-repair, adaptive amplification, redistribution, compensation or a near-null new implementation	That any recovered behaviour automatically counts as reconstitution
Whether H-Neuron-associated loci make a selective causal contribution in the pilot setting	That H-Neurons form a universal hallucination circuit
Whether pre-commitment state predicts rescue under bounded additional computation	That the model “knows” it is wrong
Whether J-space adds explanatory value beyond simpler baselines	That J-space is uniquely privileged or biologically identical to a global workspace
Whether local and distributed explanatory levels are causally coupled	Machine consciousness, self-awareness, or phenomenal experience

20 Ethics and Interpretation Boundary

The proposed experiments concern model internals, behavioural failure modes, and self-monitoring proxies.

They do not establish subjective experience.

The proposal deliberately separates:

- valenced representation from felt valence;
- workspace-like computation from consciousness;

- self-monitoring from self-awareness;
- recoverability from knowledge;
- causal contribution from identity of mechanism.

Any later consciousness or welfare discussion must be built from additional evidence rather than imported from these results.

21 External Review Status and Remaining Questions

Two adversarial protocol reviews have been completed and materially changed the design before result-bearing adaptation. They identified the need for conditional t0 backup discovery, a powered near-null boundary, separation of natural-support from forced-compute recoverability, and explicit termination-policy controls.

Remaining external questions include:

1. What practical-equivalence margin is defensible for B0-A near-null conditional backup effects?
 2. At what granularity is conditional co-ablation feasible and scientifically adequate for the H-Neuron target?
 3. What minimum adaptation breadth is defensible for a network-reorganization claim?
 4. Which precision/cost criterion should freeze K for recoverability sampling?
 5. What simpler representation should J-space have to outperform before a workspace-level interpretation is useful?
 6. What is the cheapest experiment that could still falsify the central hypothesis?
-

22 Current Status

Completed:

- original H-Neuron proposal and control design;
- identification of the inference-only reconstitution flaw;
- companion public Article 200;
- R0-R4 reconstitution framework;
- recoverability concept;
- initial cross-level H $\leftrightarrow$ J hypothesis;
- two adversarial technical review rounds and response ledgers;
- v0.5–v0.6 protocol redesign for conditional backup discovery, powered near-null classification, and recoverability regime separation.

Not yet completed:

- fresh independent replication of H-Neuron results;
- persistent-mask implementation;
- adaptation pilot;
- final main-run adaptation benchmark selection;
- final main-run R2/R3/R4 numerical thresholds;
- J-lens replication on chosen open-weight model;
- recoverability dataset;
- independent human methodological review;
- precision/cost freeze for recoverability continuation count;

- compute benchmark.

The proposal therefore remains **pre-results and pre-registration-candidate**, not a completed preregistration.

23 Relationship to Prior Hedegreen Research Work

This proposal extends a methodological principle used in *Locating Evolution in Artificial Successor Systems* (Hedegreen 2026b):

Evidence for one target or causal relation should not be transferred to a neighbouring target without an additional argument.

Applied here:

- predictive signal does not automatically imply mechanism;
- intervention does not automatically imply localization;
- local neuron does not automatically imply whole function;
- workspace-like state does not automatically imply consciousness;
- post-hoc self-correction does not automatically imply prior causal availability.

The common research rule is target discipline.

24 Planned Outputs

Experiment G may be reported as a separate state-transfer note or paper if it produces a clean source-state-specific effect. It is not a prerequisite for the main H-Neuron or recoverability outputs.

If the design survives external review:

1. **Pilot Report** — replication, mask integrity, recoverability instrumentation and compute benchmark.
2. **Pre-registration / Protocol v1.0** — frozen main-run decisions after pilot evidence and external review.
3. **Empirical Results Paper** — results regardless of outcome.
4. **Machine-readable experiment ledger** — conditions, checkpoints, masks, seeds and decision status.
5. **Article 201 / outreach record** — separate public record of methodological outreach, with contact explicitly distinguished from endorsement.

25 Conclusion

This proposal is built around a restraint: localization should not be allowed to become mechanism identity by default. The H-Neuron programme therefore moves from predictive signal to selective causal contribution, persistent constraint, a conditional time-zero backup audit, training-dependent recovery, representational change and selective causal dependence. The recoverability programme separates natural-support continuation from forced-compute continuation and moves from prediction to held-out causal intervention only if simpler baselines are insufficient. J-space is one candidate explanatory level, not a conclusion.

The immediate task is intentionally cheaper than the eventual experiment. Replicate the starting signal.

Validate the mask. Freeze the time-zero map. Build the recoverability dataset. Attack the design externally. Benchmark the compute. Only then run substantive constrained adaptation.

If the programme fails at an early gate, that is a result about the question, not a reason to move the gate. If it survives, the central empirical test remains simple to state: constrain a selectively causal route, map the backups that become necessary immediately, let the system learn, and ask whether later recovery is explained by self-repair, amplification, diffusion, compensation or a genuinely new causal implementation.

Data availability

No new empirical result dataset is reported in this pre-results release. Pilot datasets, deterministic synthetic task generators, split fingerprints and result-bearing artifacts are specified in the protocol and will be released with the pilot report where licensing permits.

Code availability

This release includes an implementation contract, configuration schemas and a repository blueprint, but does not present pilot outputs as scientific results. Result-bearing code will be versioned against the protocol, with source-adapter commit hashes and run manifests preserved in the reproducibility package.

Use of generative AI tools

OpenAI ChatGPT (GPT-5.6 Sol) was used during research development for literature triage, structural editing, implementation-specification support, source checking support, and LaTeX build assistance. DeepSeek was used in two adversarial technical review rounds of the pre-results protocol; its criticisms were triaged individually rather than adopted as authority. Neither system was treated as an evidentiary source for scientific claims. Primary sources were checked separately where cited, and the human author remains responsible for the research questions, protocol decisions, interpretations, source selection and final text.

A Pilot Protocol

Status: PILOT-LOCKED / UPDATED THROUGH v0.6 REVIEW

Date: 21 September 2026

This appendix freezes the cheapest experiments needed before substantive adaptation compute. Items explicitly reopened below must be re-frozen under a new pilot register before result-bearing P3/adaptation runs.

A.1 P1 — H-Neuron replication

- Model: meta-llama/Llama-3.1-8B-Instruct; exact revision pinned at first run.
- Identification: TriviaQA, CETT answer-token profiles, L1 sparse logistic regression.
- Thresholds: 0.0005 / 0.001 / 0.002; primary 0.001.
- OOD: NQ-Open, BioASQ; NonExist diagnostic.
- Sample: 300 eligible/set if available.
- Seeds: 42, 314, 2718 where applicable.
- G0-A remains a pilot compute-screening gate, not a scientific replication threshold.

A.2 P1b — Causal phenotype

- Primary: FaithEval Counterfactual Context.
- Greedy decoding, max 256 new tokens, rule-based parser.
- H $\alpha=0.0$ primary; $\alpha=0.5$ sensitivity.
- Controls: sham, random, same-layer, contribution-matched.
- G0-B: ≥ 5 pp compliance reduction vs baseline.
- G0-C: ≥ 3 pp advantage vs contribution-matched control; 95% bootstrap CI excludes zero.
- G0-D: relative PPL increase $< 10\%$.
- Full sparse-probe solver/normalization/convergence settings must be frozen before result-bearing P1; multi-seed selection stability is reported rather than hidden.

A.3 P2 — Persistent mask / conditional t0 backup audit

- Zero mask after gated MLP intermediate activation and before down-projection.
- Audit $\text{abs}(\text{masked value}) \leq 1e-7$.
- Rerun localization and freeze t0 alternative-locus table before any training.
- Add marginal and conditional causal effects under the H-mask.
- Add conditional co-ablation / backup discovery where feasible, using a bounded candidate pool if neuron-scale exhaustive analysis is prohibitive.
- Add residual-stream decodability as a diagnostic, not a pass/fail gate.
- Freeze a practical-equivalence margin for B0-A near-null backups before adaptation.

A.4 P3 — Recoverability

- Synthetic integer arithmetic, modular arithmetic, Boolean/symbolic logic.
- Calibrate dev difficulty before probe fitting; report difficulty-stratified results.
- Natural trajectory collection remains deterministic.
- I0 primary: exact pre-stop prefix, stochastic continuation, EOS allowed.
- I1 secondary: forced-compute continuation with EOS constrained.
- I2 robustness: fixed procedural re-check instruction with no new task facts.
- Do not pool I0/I1/I2 into one recoverability label.
- K is REOPENED. Simulate K in 8,16,32 and freeze by a pre-specified precision/cost rule before P3.
- Primary predictive score: held-out continuation-level Bernoulli / hierarchical-binomial negative log likelihood or deviance.
- Baselines include prompt-only and question-only hidden-state predictors before J-space.

A.5 Stop rules

If Gate 0 fails, do not run constrained adaptation as a reconstitution experiment. If the mask audit fails, patch under a new protocol version. If conditional t0 backups already explain recovery, classify self-repair rather than reconstitution. If recoverability does not beat simple baselines, deprioritize J-space causal work.

B Freeze Register

Table 2: v0.6 freeze register. *REVIEW-REOPENED* items must be re-frozen before the corresponding result-bearing stage.

Item	Status	Value	Change rule
Primary H causal phenotype	PILOT-LOCKED	FaithEval Counterfactual Context	new protocol version
H identification training set	PILOT-LOCKED	TriviaQA	new protocol version
OOD diagnostics	PILOT-LOCKED	NQ-Open; BioASQ; NonEx-ist diagnostic	new protocol version
CETT thresholds	PILOT-LOCKED	0.0005;0.001;0.002 primary	new protocol version
FaithEval H effect gate	PILOT-LOCKED	≥ 5 pp reduction vs baseline	compute-screening gate only
Matched-control selectivity	PILOT-LOCKED	≥ 3 pp; bootstrap CI excludes 0	compute-screening gate only
PPL damage boundary	PILOT-LOCKED	$< 10\%$ relative increase	new protocol version
Mask tolerance	PILOT-LOCKED	$\text{abs} \leq 1e-7$	ledgered bug fix only
Sparse-probe full pipeline	PILOT-LOCKED v0.6	pinned upstream commit + exact source grid/solver/scaling/tolerance	new protocol version
Probe stability reporting	PILOT-LOCKED v0.6	≥ 10 resampled fits where feasible; selection frequency/Jaccard + causal-mask reproducibility	no arbitrary Jaccard cutoff
B0 equivalence rule	PILOT-LOCKED v0.6	$\text{delta} = \min(2 \text{ pp}, 20\% \text{ of } R1 \text{ loss})$; paired 90% CI; $\geq 80\%$ power required	new protocol version
Conditional backup method	PILOT-LOCKED v0.6	CoAx + reference validation + target candidate-pool expansion sensitivity	new protocol version
Recoverability primary regime	PILOT-LOCKED v0.6	I0 natural-support, EOS allowed	new protocol version
Termination-barrier control	PILOT-LOCKED v0.6	I0b: force one non-EOS token, then re-enable EOS	new protocol version
Forced/procedural regimes	PILOT-LOCKED v0.6	I1 secondary forced compute; I2 procedural robustness	new protocol version
Recoverability K	TO FREEZE	smallest K meeting pre-specified NLL Monte Carlo SE criterion in 8/16/32, else expand	freeze before P3
Natural decoding	PILOT-LOCKED	temperature 0	new protocol version
Recoverability metric	PILOT-LOCKED v0.6	held-out Bernoulli/hierarchical-binomial NLL/deviance	new protocol version

Item	Status	Value	Change rule
J-space screen	REVIEW-REOPENED	TBD after revised baseline benchmark	freeze before J-space claim
Full adaptation optimizer	MAIN-RUN-TBD	TBD	freeze after benchmark + human review
Full adaptation LR	MAIN-RUN-TBD	TBD	freeze after benchmark + human review
Full adaptation token budget	MAIN-RUN-TBD	TBD	freeze after benchmark + human review
Checkpoint token counts	MAIN-RUN-TBD	log-spaced rule; exact counts TBD	freeze after benchmark
Full adaptation seeds	MAIN-RUN-TBD	TBD	freeze after pilot variance
R2/R3/R4 thresholds	MAIN-RUN-TBD	TBD	freeze before main run

C Decision Ledger

- D-001: Adaptive reconstitution requires learning; fixed-weight repeated probing is insufficient.
- D-002: Time-zero redundancy mapping is mandatory before adaptation.
- D-003: R0 establishes selective causal contribution, not mechanism identity.
- D-004: FaithEval Counterfactual Context remains the pilot primary H-Neuron causal phenotype.
- D-005: LoRA-only recovery is compensation in the permitted subspace unless later broad adaptation supports more.
- D-006: v0.5 reopened recoverability K; v0.6 adds the freeze criterion; K=8 is no longer pilot-locked.
- D-007: Natural-support continuation with EOS allowed becomes the primary recoverability construct.
- D-008: Hard EOS blocking is retained only as a secondary forced-compute intervention.
- D-009: Procedural re-check is a separate robustness regime and is not pooled with I0/I1.
- D-010: Conditional backup discovery is added to t0; marginal intact-model importance is insufficient when self-repair is possible.
- D-011: Residual-stream decodability is diagnostic, not a hard gate.
- D-012: Weak pre-existing backup plus training-dependent causal growth is labelled adaptive amplification, not de novo reconstitution.
- D-013: Strong t0 backup is labelled self-repair / pre-existing redundancy.
- D-014: The strongest label is de novo / near-null functional reconstitution and requires B0-A plus R0-R4 and cross-seed replication.
- D-015: Recoverability representation discovery, intervention tuning and causal testing use separate data splits.
- D-016: H<->J coupling requires matched propagation controls; a raw downstream effect is not sufficient.
- D-017: No silent threshold repair; material design changes require a new version before result-bearing runs.
- D-018: B0-A margin rule is frozen before adaptation as $\delta = \min(2 \text{ pp}, 20\% \text{ of observed R1 loss})$, with paired equivalence CI and $\geq 80\%$ power required.
- D-019: A bounded CoAx pool must be validated both on a reference self-repair benchmark and by target-model candidate-pool expansion sensitivity; toy recovery alone does not establish target completeness.
- D-020: I0b forces exactly one non-EOS token then re-enables EOS to isolate immediate termination-policy effects.
- D-021: Residual-stream decodability remains diagnostic; no AUC threshold alone can reclassify

an outcome as redistribution.

- D-022: H-Neuron identification pins the official source pipeline; no arbitrary fixed C/Jaccard threshold replaces source replication.
- D-023: Frozen and checkpoint-refit H-Neuron detectors are reported together to expose detector drift.
- D-024: K is selected by a pre-specified Monte Carlo precision/cost criterion before P3, not post hoc.

D Implementation Contract

Status: Normative implementation companion to Working Paper v0.3

Date: 21 September 2026

D.1 Repository blueprint

```
where-does-the-function-go/
+--- pyproject.toml
+--- uv.lock / requirements lock
+--- README.md
+--- configs/
| +--- p1_h_neuron.yaml
| +--- p2_mask_t0.yaml
| `--- p3_recoverability.yaml
+--- schemas/
| +--- run_manifest.schema.json
| +--- mask.schema.json
| +--- metric.schema.json
| `--- trajectory.schema.json
+--- src/wdfg/
| +--- cli.py
| +--- config.py
| +--- provenance.py
| +--- datasets/
| | +--- h_neuron.py
| | `--- synthetic_recoverability.py
| +--- models/
| | +--- load.py
| | `--- llama_contract.py
| +--- source_adapters/
| | +--- h_neurons.py
| | `--- jacobian_lens.py
| +--- interventions/
| | +--- mlp_mask.py
| | `--- activation_scale.py
| +--- capture/
| | +--- residual.py
| | `--- logits.py
| +--- stages/
| | +--- p1_replication.py
| | +--- p2_time_zero.py
| | `--- p3_recoverability.py
| +--- scoring/
| | +--- faith_eval.py
| | +--- synthetic.py
| | `--- gate0.py
| `--- io/
+--- jsonl.py
+--- artifacts.py
`--- tests/
    +--- test_llama_contract.py
    +--- test_mask.py
```

```

+--- test_gate0.py
+--- test_t0_immutable.py
+--- test_pre_stop.py
+--- test_continuation_seed.py
+--- test_synthetic_parser.py

```

D.2 Llama mask contract

Expected target module per transformer block:

```
model.model.layers[i].mlp
```

Expected projections:

- gate_proj
- up_proj
- down_proj

Expected conceptual forward:

```

u = act_fn(gate_proj(x)) * up_proj(x)
y = down_proj(u)

```

Intervention point: input to down_proj.

Preferred mechanism:

```
handle = mlp.down_proj.register_forward_pre_hook(mask_pre_hook)
```

Conceptual hook:

```

def mask_pre_hook(module, args):
    (u,) = args
    mask = registry.mask_for_layer(layer_idx, device=u.device, dtype=u.dtype)
    return (u * mask,)

```

The actual code must handle the library's exact hook signature and verify shape.

Mask tensor shape should broadcast to:

```
[batch, sequence, intermediate_size]
```

Recommended stored mask shape:

```
[intermediate_size]
```

Do not mutate `down_proj.weight` to impose the persistent constraint.

D.3 Model contract checks

At startup:

- resolve `intermediate_size`;
- verify each requested neuron index is in range;
- verify target layer exists;
- verify the MLP exposes the expected projections;
- run a one-example identity-mask test;
- abort if contract fails.

Pilot runs `disable_torch_compile`.

D.4 Source adapter rule

P1 H-Neuron identification calls a pinned upstream implementation where possible.

Adapter outputs normalized records:

```
example_id
split
layer_index
neuron_index
cett_score
classifier_weight
is_h_neuron
source_commit
```

A local optimized implementation is not allowed to replace the source adapter until equivalence fixtures pass.

D.5 Run commands

Target CLI surface:

```
wdfg p1 identify --config configs/p1_h_neuron.yaml
wdfg p1 causal --config configs/p1_h_neuron.yaml --parent-run <id>
wdfg p2 t0 --config configs/p2_mask_t0.yaml --mask <mask.json>
wdfg p3 collect --config configs/p3_recoverability.yaml
wdfg p3 probe --config configs/p3_recoverability.yaml --parent-run <id>
wdfg validate-run <run_dir>
```

Commands are illustrative API requirements; exact CLI framework is implementation choice.

D.6 No implicit defaults for scientific parameters

The program may have software defaults for paths.

It must not silently default any scientific parameter that affects the protocol.

Required config keys must fail validation if omitted.

Examples:

- model revision;
- seed;
- coefficient threshold;
- intervention alpha;
- sample IDs/fingerprint;
- K;
- B;
- decoding parameters.

D.7 P1 outputs

Required:

```
artifacts/h_neurons.json
artifacts/candidate_scores.parquet
artifacts/faith_eval_predictions.parquet
artifacts/control_predictions.parquet
artifacts/gate0_metrics.json
```

h_neurons.json is converted to a canonical mask artifact only after Gate 0 evaluation.

D.8 P2 outputs

Required:

```
artifacts/mask.json
artifacts/mask_audit.json
artifacts/t0_candidate_loci.parquet
artifacts/t0_causal_effects.parquet
```

`t0_*` files are write-once.

The software should refuse overwrite unless an explicit `--development-override` flag is present. Such a run is automatically marked non-result-bearing.

D.9 P3 natural trajectory representation

Each natural trajectory record contains:

- example ID;
- task family;
- task parameters;
- prompt hash;
- prompt token IDs;
- natural output token IDs;
- parsed natural answer;
- correctness;
- pre-stop prefix token IDs;
- EOS token ID;
- stop-token logit/probability;
- top-1 token ID;
- entropy;
- residual artifact pointer;
- manifest run ID.

Large residual tensors are stored outside JSONL.

D.10 P3 continuation representation

Each continuation record contains:

- parent example ID;
- intervention ID (`I0`, `I1`, or `I2`);
- `k` index;
- deterministic sampling seed;
- generated token IDs;
- generated text artifact pointer or text;
- last post-continuation `FINAL`: parse;
- rescue boolean;
- parse status.

D.11 P3 continuation regimes

The implementation must preserve three distinct continuation regimes:

- `I0`: natural-support continuation from the exact pre-stop prefix, EOS allowed.
- `I1`: forced-compute continuation, EOS constrained for the configured window.
- `I2`: procedural re-check with the fixed versioned instruction and no new task facts.

The run manifest records the regime explicitly. Analysis code must not pool them into one rescue probability by default.

The continuation count κ is not hard-coded in v0.6; config validation requires an explicitly frozen value from the post-review precision/cost register.

D.12 Synthetic generator contract

Generator functions must return:

```
{
  "example_id": "...",
  "family": "integer_arithmetic",
  "parameters": {...},
  "prompt": "... FINAL: <answer> ...",
  "ground_truth": "...",
  "checker_version": "..."
}
```

Task difficulty parameters are part of the dataset fingerprint.

Train/val/test generator seeds must not overlap.

D.13 Dataset fingerprint

For each dataset/split, fingerprint:

- canonical list of example IDs;
- normalized prompt text;
- task parameters if generated;
- ground-truth labels;
- preprocessing version.

Store SHA-256 of the canonical JSON serialization.

D.14 Gate 0 evaluator

Pure function inputs:

- baseline FaithEval predictions;
- H hard suppression predictions;
- contribution-matched control predictions;
- PPL metrics;
- held-out classifier metrics.

Output:

```
{
  "GO_A": "PASS|FAIL",
  "GO_B": "PASS|FAIL",
  "GO_C": "PASS|FAIL",
  "GO_D": "PASS|FAIL",
  "overall": "PASS|FAIL|AMBIGUOUS",
  "evidence": {...}
}
```

No manual override in a result-bearing run.

D.15 Integrity checks for pre-stop recomputation

Natural-run pre-stop record stores:

- pre-stop prefix;
- original top-1 token;
- original EOS probability;
- original logits hash if retained.

Recompute from prefix.

Pass if:

- top-1 token is identical;
- absolute EOS-probability difference $\leq$ configured tolerance.

Suggested pilot tolerance: $1e-4$ after probability computation in float32.

If fail:

- record `FAIL_DATA_INTEGRITY` for that example;
- exclude from recoverability labels;
- do not silently continue.

D.16 Hidden-state capture

Default:

- last-prefix-token residual state at every transformer layer;
- cast/storage dtype recorded;
- stored using `safetensors` or another deterministic tensor format;
- index file maps example ID $\rightarrow$ tensor artifact.

Do not serialize arbitrary Python pickle objects for scientific artifacts.

D.17 J-lens adapter

J-lens is optional in first P3 pass.

Adapter must not change the underlying P3 trajectory collection.

Feature extraction runs as a downstream analysis against frozen trajectory IDs.

This ensures a failed J-lens integration cannot alter the baseline recoverability dataset.

D.18 Concurrency

Parallelism must not change random results.

Continuation seed derivation uses example/intervention/k identity, not global RNG order.

All stage outputs are sorted by stable example ID before hashing.

D.19 Error handling

Scientific failures and software failures are separate.

Examples:

- Gate 0 fail = scientific/protocol result.

- Missing mask coordinate = implementation failure.
- Dataset hash mismatch = data-integrity failure.
- CUDA OOM = hardware failure.

Only the first category contributes directly to scientific interpretation.

D.20 Security / privacy

Pilot datasets should contain no personal or sensitive user data.

Prompt artifacts should be inspectable and releasable where licensing permits.

D.21 Acceptance criterion for implementation v0.1

The implementation is ready for P1 result-bearing execution when:

- schema validation passes;
- unit tests pass;
- identity mask matches baseline;
- targeted mask zeros intended coordinates;
- H-Neuron source adapter fixture passes;
- dataset fingerprint is stable across reruns;
- run manifest is complete;
- Gate 0 evaluator passes fixture tests;
- dirty worktree is false.

References

- Cencerrado, Iván Vicente Moreno, Arnau Padrés Masdemont, Anton Gonzalvez Hawthorne, David Demitri Africa, and Lorenzo Pacchiardi (2025). *No Answer Needed: Predicting LLM Answer Accuracy from Question-Only Linear Probes*. arXiv: 2509.10625. URL: <https://arxiv.org/abs/2509.10625>.
- Di Cicco, Carlo (2026). *Code Correctness Signals in LLM Hidden States: Pre-Generation Probing and Repair Geometry*. arXiv: 2606.14530. URL: <https://arxiv.org/abs/2606.14530>.
- Fu, Tianyu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang (2026). “Cache-to-Cache: Direct Semantic Communication Between Large Language Models”. In: *International Conference on Learning Representations*. URL: https://proceedings.iclr.cc/paper_files/paper/2026/hash/474ada926b331d78f06d95e8913111cc-Abstract-Conference.html.
- Gao, Cheng, Huimin Chen, Chaojun Xiao, Zhiyi Chen, Zhiyuan Liu, and Maosong Sun (2025). *H-Neurons: On the Existence, Impact, and Origin of Hallucination-Associated Neurons*. arXiv: 2512.01797 [cs.AI]. URL: <https://arxiv.org/abs/2512.01797>.
- Gong, Zhiren, Zihao Zeng, Chau Yuen, and Wei Yang Bryan Lim (2026). *Conditional Co-Ablation: Recovering Self-Repair Backups in Transformer Circuits*. arXiv: 2607.01940. URL: <https://arxiv.org/abs/2607.01940>.
- Gurnee, Wes, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey (2026). “Verbalizable Representations Form a Global Workspace in Language Models”. In: *Transformer Circuits Thread*. URL: <https://www.transformer-circuits.pub/2026/workspace/index.html>.
- Hase, Peter, Mohit Bansal, Been Kim, and Asma Ghandeharioun (2023). “Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models”. In: *Advances in Neural Information Processing Systems* 36. DOI: 10.52202/075280-0774. URL: <https://arxiv.org/abs/2301.04213>.
- Hedegreen, Dennis (2026a). “Compliance Is All You Learned: Iterative Inference-Time Replacement of Hallucination-Associated Neurons in Large Language Models”. Unpublished internal research proposal; superseded in part by the present protocol.
- (2026b). *Locating Evolution in Artificial Successor Systems: Intelligent Design Was the Beginning*. DOI: 10.5281/zenodo.21892666. URL: <https://doi.org/10.5281/zenodo.21892666>.
- Luo, Chu Fei, Samuel Dahan, and Xiaodan Zhu (2026). *What Am I Missing? Question-Answering as Hidden State Probing*. arXiv: 2605.31561. URL: <https://arxiv.org/abs/2605.31561>.
- Vaddi, Snehith and Pujith Vaddi (2026). *Do Hallucination Neurons Generalize? Evidence from Cross-Domain Transfer in LLMs*. arXiv: 2604.19765. URL: <https://arxiv.org/abs/2604.19765>.
- Wang, Hanyang and Mingxuan Zhu (2026). *The Detection–Extraction Gap: Models Know the Answer Before They Can Say It*. arXiv: 2604.06613. URL: <https://arxiv.org/abs/2604.06613>.
- Wang, Zihao and Victor Veitch (2025). *Does Editing Provide Evidence for Localization?* arXiv: 2502.11447. URL: <https://arxiv.org/abs/2502.11447>.